基于大数据的媒体传播分析及影响力评估应用创新

摘 要：如何追溯新闻报道在融媒体、大数据环境下的全球实时落地采用及传播情况，有效评估新闻报道的综合影响力、辅助采编决策，是媒体融合发展面临的新课题。通过对跨媒体大数据融合技术、内容智能比对技术、跨平台传播链路分析技术、媒体传播影响力评估技术等技术上的创新，实现融合媒体报道在全球的实时落地采用、传播分析和综合影响力评估，构建一套媒体传播分析及影响力评估应用体系，将有助于指导新闻媒体行业的采编决策，增强融合报道、对外报道传播能力，提升媒体影响力。

关键词：大数据；采用分析；媒体传播分析；影响力评估

中图分类号：G206　　文献标识码：A

文章编号：1671-0134（2017）10-122-03　　DOI：10.19483/j.cnki.11-4653/n.2017.10.051

文 / 陈 珺 陈辛夷 苏 宇

引言

近年来，新媒体平台的飞速发展，使得新闻报道发布和传播渠道呈现多元化，从单一的平面媒体向多元化的新媒体及全媒体转变，一条新闻报道往往会在报纸报刊、新闻网站、两微一端和海外社交媒体等多种平台和媒体形态上发布传播，如何通过大数据技术和智能分析算法全面实时掌握新闻报道在跨平台、多种媒体形态的落地采用和传播影响力情况，辅助于采编决策，是媒体融合发展面临的新课题和现实需求。

通过利用新兴的大数据智能分析技术，及时搜集处理互联网海量信息，精确定位新闻报道在媒体上的落地采用信息，跨渠道多维度分析新闻传播效果，评估稿件、专题、产品的综合影响力，快速编制分析报告，是新闻信息生产全流程不可或缺的重要环节，是构建大数据驱动采编和传播决策的重要组成部分，对于新闻媒体的传播能力建设具有重要的意义。

1. 难题与挑战

要实现基于大数据的实时、自动和全面的新闻报道信息传播分析和影响力评估，面临着诸多的技术难题与挑战。

挑战一：如何利用分布式云计算技术，及时准确地获取互联网多来源、海量、异构且动态更新的媒体数据信息。融媒体传播分析需要全媒体数据，既需要获取电子报刊类传统媒体数据，也需要获取新闻网站、“两微一端”和海外社交媒体平台的数据。要实现自动、实时监测和采集各类平台，各种媒体形态的海量互联网信息，每天需要解析、清洗、处理数百万条到数千万条异构原始数据，构建多来源、海量和动态的融合媒体大数据平台存在很大的技术挑战。

挑战二：如何利用大数据分析与处理技术，实现自动、智能、及时、准确的新闻报道落地采用分析计算。现在的新闻报道种类丰富，包括文字、图片图表、视频、多媒体等多种类型以及多种语种，想要从千差万别的海量异构媒体信息中及时识别和准确定位采用，需要设计和不断调整优化不同类型报道的采用判定算法，能够更加智能处理各类复杂情况，这个对分析技术是一个很大的挑战。

挑战三：如何利用报道内容智能化关联分析技术，链接传统媒体与新媒体的平台传播鸿沟，实现跨媒体平台的内容传播分析。全媒体时代，报道信息跨媒体平台进行传播，需要通过报道内容智能化分析技术实现信息关联和分析，全面掌握报道信息的全媒体平台传播情况。

挑战四：如何设计科学的传播影响力评估指标和评价体系，量化评价传播贡献。如何使媒体传播影响力的测定更加科学、理性、全面、规范，如何建立一套科学合理的量化指标，实现新闻报道的传播影响力定量分析和评价，达到生产“苦劳”和影响“功劳”的综合评价目的。

2. 总体架构

一套基于大数据的媒体传播分析及影响力评估应用体系由“一个平台、一套知识库、七个技术层级、六大应用功能”组成，如图 1 所示。

2.1 “一个平台 + 一套媒体知识库”——媒体大数据平台

媒体大数据平台采用先进的大数据框架体系构建，融合了传统媒体（即新闻网站和电子报纸）、“两微一端”（即微博、公众微信号和移动新闻客户端）、海外社交媒体平台（包括脸谱、推特和优兔）等多种类型的媒体数据，并积累了大量的媒体基础信息，建立描述全球媒体属性的媒体资料信息库。

2.2 七个技术层级

包括：数据总线层、数据采集层，数据接入层，数据整合层，数据资源层，业务分析层和应用服务层。

数据总线层：实现整个平台的底层硬件、数据资源、技术组件和功能应用的通信链路。基于服务总线技术，解决多数据、多模块、多应用间的协同、共享、通信和管理，实现数据、模块、应用的服务化注册、管理和调用以及数据、组件和应用的服务化。

数据采集层：实现各渠道采集数据的统一采集管理，利用分布式云采集技术，确保数据采集的及时性，构建基础网络设施，确保网络的高可用性以及代理资源的高可靠性。

数据接入层：实现新闻报道数据、互联网新闻数据、“两微一端”数据、社交媒体数据等的接入。

数据整合层：实现多源异构数据的抽取、转化和融合。

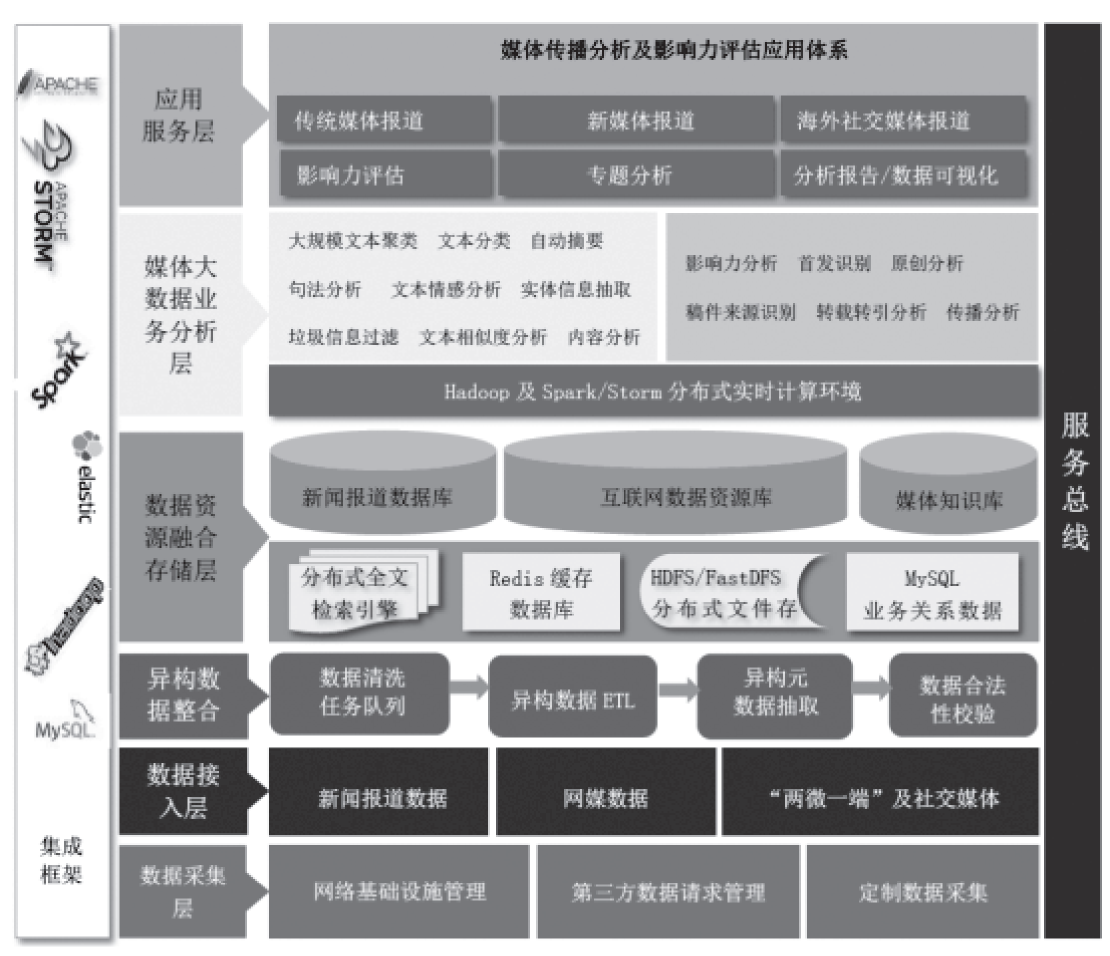


图 1　系统总体架构图

搭建分布式数据处理任务队列，实现海量任务的 ETL 过程，对数据进行初步的结构化处理；针对异构数据，进行精确字段抽取及格式化；对结构化数据进行合法性校验，过滤垃圾无用信息，对于不合法信息进行日志记录，有效数据则提交存储；根据源数据结构将内容解析出来，并进行特殊字段的转换。

数据资源层：实现异构数据资源的融合存储管理。处理后的结构化数据将根据不同的数据类型加载到数据资源层的关系型数据库、全文检索数据库、分布式文件系统上。

业务分析层：实现内容智能分析、采用分析、传播分析、影响力评估算法、数据统计等具体的数据分析处理。面对海量互联网媒体数据，为了高效、实时进行数据处理，需要构建一套高性能分布式实时计算环境，采用 Hadoop+Spark 的分布式计算框架可以最大限度地发挥软硬件资源的计算能力。在此基础上，构建“两套分析引擎”，其中：文本智能分析算法引擎负责自然语言处理领域的核心算法支持，包括：大规模文本聚类、实体信息抽取、句法分析、文本分类、自动摘要、垃圾信息过滤和文本情感分析等。传播分析及影响力分析引擎负责业务层面的模型算法实现，包括：原创分析、稿件来源识别、首发识别、转载转引识别、传播分析、影响力指标体系构建模型等。

应用服务层：实现各类具体的功能模块和集成应用服务，并提供对外数据服务。采用基于数据总线的分布式微服务集群架构，可以应对当业务压力上升，服务器容量难以评估，小服务资源浪费的问题，提高集群利用率。同时，能够提高 IT 架构的灵活性，快速响应业务环境变化及内部需求对业务流程优化提出的要求，最大限度复用现有 IT 资源，避免重复构建。

2.3 六大应用功能

传统媒体报道分析应用：实现新闻报道在纸媒、网媒、微信、客户端上的全球跨平台实时采用数据及互动数据监测分析。

新媒体报道分析应用：实现微博微信账号的实时监测分析，包括对账号粉丝量、发稿量、稿件内容及互动量的监测分析。

海外社交媒体报道分析应用：实现海外主流社交平台上的帐号分析，贴文的内容分析和传播分析，与其他主流媒体账号的对比分析，被海外媒体上引用的数据监测分析等。

影响力评估指标体系：实现新闻报道的跨渠道媒体传播效果评估指标体系，由全网影响力指数和不同渠道影响力指数构成。针对每个传播渠道，各指标体系涵盖阅读、互动、采用三个评估粒度。

专题分析应用：实现针对重大专题报道的事件分析、报道分析和影响力分析。深度分析重大专题事件报道在全球媒体的传播效果以及和同业媒体报道的传播对比分析，分析该事件的发展趋势、焦点脉络、主要观点、媒体和网民关注情况、舆论情感发展趋势，实现专题报道的影响力评估。

分析报告和数据可视化：实现新闻报道的多维度分析

报告的自动生成，能够提供丰富的数据可视化展现。

3. 关键技术创新

3.1 跨媒体大数据获取与融合技术

利用分布式云采集技术、多源媒体数据融合技术和媒体数据云服务技术，构建媒体大数据平台，自动采集引进全球媒体网站、电子报纸、新媒体数据、海外社交媒体数据等，提升了站点覆盖面、数据规模和更新速度。

部署了可采集全球中英文网站的分布式云采集端，通过本地采集和境外部署回传数据。

实现自采数据和多个第三方数据的同步整合，融合和利用多方数据资源，形成多个数据云之上的“集合云”。

覆盖广泛的媒体数据类型，实现新闻、电子报纸、两微一端和海外社交媒体等多种数据类型的实时动态采集。

通过数据融合处理，面向各类应用系统提供基础数据云服务。

3.2 基于流式处理的稿件采用智能比对技术

采用领先的文本语义分析技术，结合业务规则，研发了具有自主知识产权的稿件采用智能比对技术，能够准确定位新闻报道稿件在中英文媒体中的落地采用情况。

通过总结业务经验和业务规则，并结合机器学习模型，不断优化和修正稿件准确度判定参数，形成了业务认可的稿件采用判定规则和阈值设置。

利用算法实现了文字和图片报道的自动采用计算，不仅可以识别稿件显性采用（标注显性关键词的采用），还可以识别稿件隐形采用（未标注显性关键词的采用）。

通过引入新兴的 Spark 大数据流式处理框架，实现分钟级的采用结果更新速度，可提供近实时的采用分析数据。

实现了英文稿件的自动采用比对和传播分析。

3.3 跨平台传播链路分析技术

融媒体传播时代，稿件的传播往往跨越多个媒体平台进行传播，而稿件在不同平台之间的传播关系往往难以关联和发现。通过研发内容智能化关联分析技术，解决传统媒体与新媒体传播链路识别问题，实现了单篇稿件的跨媒体采用落地。

积累主流媒体知识库，自动关联同一稿件的多个发布渠道，识别多媒体平台发布产生的跨平台传播。

通过文本内容相似特征比对技术，识别不同媒体平台中的同一稿件落地采用。

基于时间特征和内容相似特征，自动识别和关联单篇稿件跨媒体平台的传播链路，追溯稿件传播过程。

3.4 媒体传播影响力评估技术

基于媒体大数据信息，基于内容传播指标和受众互动指标，形成稿件综合影响力评价模型，提供单篇稿件、线路和部门等多维度的影响力评价结果。

不同于已有的单纯依靠粉丝数、关注数等影响力评价单一指标体系，形成了数量评价指标、内容传播指标和受众互动指标的综合影响力评价指标体系。

基于模糊综合评价方法，提供影响力量化评价和计算模型，数量化稿件、线路和部门的影响力评价。

结合跨媒体传播链路分析结果，融合多个媒体平台维度的传播影响指数，提供融媒体和跨媒体影响力量化评价。

4. 应用价值

通过技术创新和应用创新，构建基于大数据的媒体传播分析及影响力评估应用体系，有助于指导新闻媒体行业的采编决策，增强融合报道对外报道传播能力，提升媒体影响力：

实现了全网跨平台的媒体监测，实现融合媒体报道在全球的实时落地采用、传播分析和综合影响力评估，形成了一套科学合理的新闻报道评价指标。

解决了对外英文报道全球实时监测和落地采用统计的难题，通过量化分析，挖掘对外英文稿件的传播特征，有目标、有侧重的进行采编选题，实现英文稿件更加精准有效传播。

加强了对新媒体报道的监测，实现传播效果的全面有效掌握，及时了解网民聚焦和互动特征，进行更有针对性的分析，有效指导采编决策。

实现对海外社交媒体运营、分析与决策辅助功能，通过对海媒账号信息、贴文信息和互动信息的监测分析，实时掌握海媒运营情况，跟踪国际主流媒体传播热点，做到知己知彼，快速响应。

实现新闻传播影响力评估指标体系，通过跨渠道、多维度的新闻传播效果分析，评估新闻报道、专题、产品在落地传播阶段的综合影响力，对比分析同业媒体的报道情况、互联网传播情况，便于报道指挥人员、采编人员、内容运营人员调整产品结构、报道资源分配、运营策略等。

为重大专题报道提供多维度的深入的数据分析、可视化展示及专题分析报告，全面掌握专题报道情况，充分体现专题报道成效，服务于采编决策。媒

参考文献

[1] 杨伟杰，戴汝为，崔霞 . 一种基于信息检索技术的网络新闻影响力分析方法 [J]. 软件学报，2009，20（9）：2397–2406.

[2] 王友忠，曾大军，郑晓龙等 . 基于复杂网络理论的互联网新闻媒体分析 [J]. 复杂系统与复杂性科学，2009，6（3）：11–20.

[3] 王君泽，曾润喜，杜洪涛 . 基于网页转载关系判别的网络舆情传播态势分析 [J]. 情报杂志，2015，34（1）：144–149.

（作者单位：新华通讯社通信技术局）